Strategic Disinformation Generation and Detection

Wenxiao Yang (UC Berkeley)
Yunfei (Jesse) Yao (CUHK)
Pengxiang Zhou (HKUST/USC)

My coauthors

Wenxiao Yang (UC Berkeley)

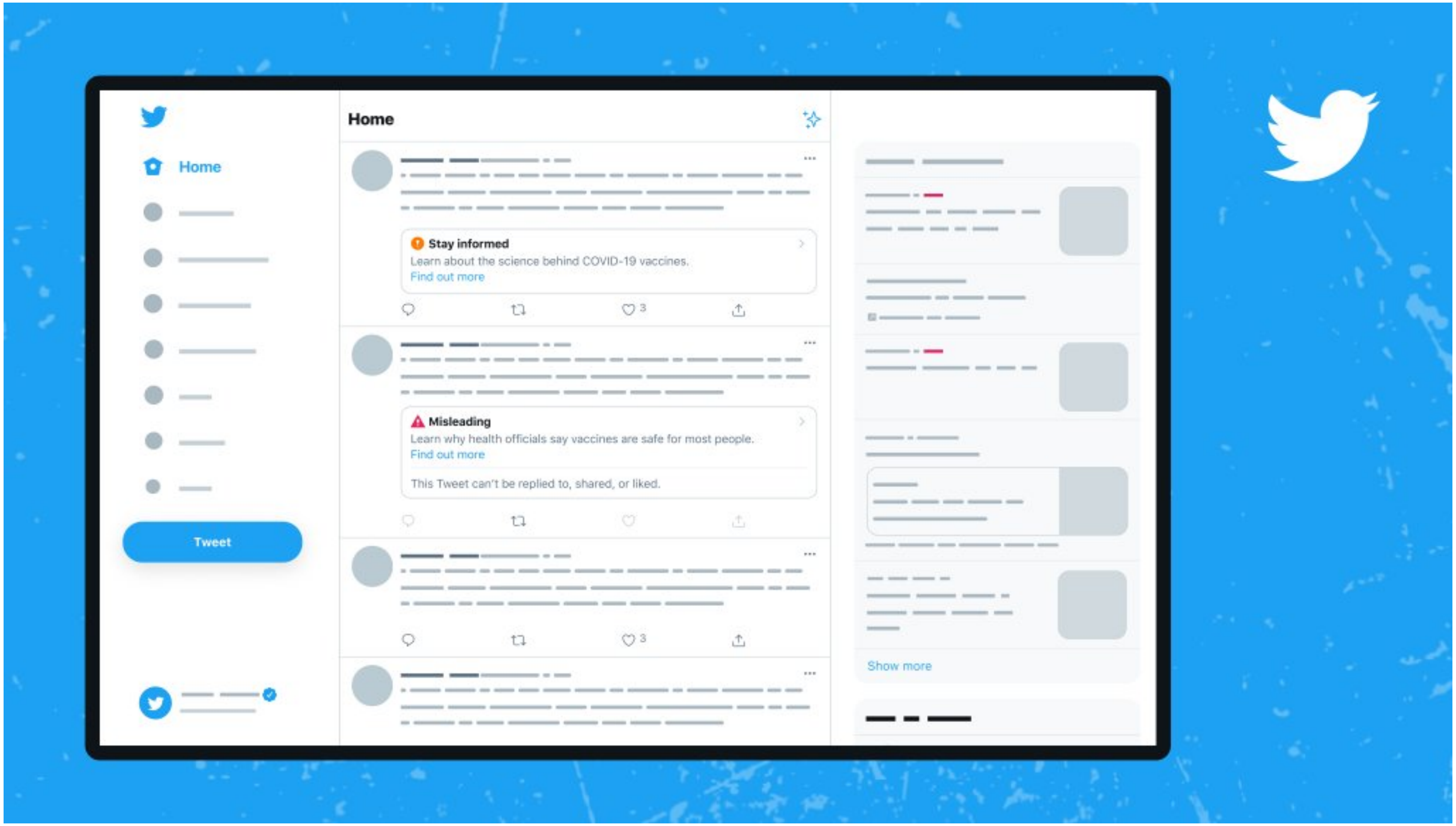
Pengxiang Zhou (HKUST/USC)

Widespread disinformation nowadays

- fake reviews
- misleading posts
- fraudulent resumes
- ad fraud

Generative AI technologies further amplified such practices

Disinformation detection (Twitter)



Disinformation detection (Yelp)

Suspicious Review Activity

We have noticed suspicious review activity for this business. This sort of activity can take many forms, including when a number of positive reviews originate from the same IP address or when we've identified reviews resulting from a possible deceptive review ring. Our automated recommendation software has taken this suspicious activity into account in choosing which reviews to display, but we wanted to call this to your attention because someone may be trying to artificially inflate the rating for this business.

Got it, thanks!

Disinformation detection

- Platforms and regulators: deploy algorithms to detect and raise warnings about disinformation
- **Detecting** disinformation remains challenging (Callander and Wilkie, 2007; Dziuda and Salas, 2018; Mattes, Popova, and Evans, 2023)

Dilemma in detecting disinformation

A. increasing the likelihood of correctly recognizing deceptive content (↑ **true-positive rate / ↓ false-negative rate**)

   cannot avoid making false-negative mistakes unless always sends an alarm

B. reducing the probability of falsely identifying genuine content as deceptive (↓ **false-positive rate**)

   cannot avoid making false-positive mistakes unless never sends an alarm

⇒ Trade-off between

Type I error (**false-positive**) & Type II error (**false-negative**)

Dilemma in detecting disinformation

- Previous work: **false negative** - a detector may fail to send an alarm when there is disinformation

  $\Rightarrow$ implicitly assuming that the false-positive rate is zero

- **False positives** are ubiquitous and economically significant:

  J.P. Morgan views false positives as a multi-billion dollar problem;

  Global business loses more than $100 billion annually due to false positives

  $\vee$

  actual fraud costs

Dilemma in detecting disinformation

- This paper: consider the possibility of **false positive** - the detector may send a false alarm without disinformation

- **Key contribution**: allow for both types of mistakes in disinformation detection.

$\Rightarrow$ qualitatively different insights about strategic communication and the design of disinformation detector

- **Other main contribution**: **endogenize** the detector design

Research questions

- How does the detection ability affect the incentive to generate disinformation?
- Given the practical constraints of classification technology, how should the detectors be designed?

Related Research

- Strategic communication:

1. verifiable disclosure (infinite cost of lying): Grossman (1981), Milgrom (1981)
2. cheap talk (zero cost of lying): Crawford and Sobel (1982)
3. costly lying (finite cost of lying): Kartik, Ottaviani, and Squintani (2007), Kartik (2009), Dziuda and Salas (2018), Balbuzanov (2019)

Recent work where firms and consumers interact: Villas-Boas, 2004; Guo and Zhao, 2009; Kuksov and Lin, 2010; Mayzlin and Shin, 2011; Sun, 2011; Zhang, 2013; Lauga, Ofek, and Katona, 2022; Chen, Du, and Lei, 2024; Ning, Shin, and Yu 2025.

Related Research

- Information design: Kamenica and Gentzkow (2011), Jerath and Ren (2021); Berman, Zhao, and Zhu (2022); Ke, Lin, and Lu (2022); Pei and Mayzlin (2022); Shin and Wang (2024); Shulman and Gu (2024)

- Strategic interactions between humans and algorithms: Liang (2019), Miklos-Thal and Tucker (2019), Salant and Cherry (2020), Calvano et al. (2020), O’Connor and Wilson (2021), Montiel Olea et al. (2022), Iyer and Ke (2024), Qian and Jain (2024), Lin, Shi, and Sun (2025)

Model

Model

- A sender (S), a receiver (R), and a detector designer

- Receiver makes a binary decision: $r_H \ vs. \ r_L$

  purchasing a product from an e-commerce seller,

  re-posting social media content,

  visiting a restaurant,

  clicking on an email link,

  sending a business contact request

- Sender: high (H) type with probability ρ, low (L) type with probability $1 - \rho$

  Sender’s private information

Model

- Sender always wants the receiver to take action r_H
- Receiver prefers to match the action with the sender's type

(sender payoff, receiver payoff)	action r_H	action r_L
type H sender	$(\Delta_H^S > 0, \Delta_H^R > 0)$	$(0, 0)$
type L sender	$(\Delta_L^S > 0, -\Delta_L^R < 0)$	$(0, 0)$

Table 1: Players' Payoffs

Model

(sender payoff, receiver payoff)	action r_H	action r_L
type H sender	$(\Delta_H^S > 0, \Delta_H^R > 0)$	$(0, 0)$
type L sender	$(\Delta_L^S > 0, -\Delta_L^R < 0)$	$(0, 0)$

- Critical belief $\hat{\rho}$: receiver is indifferent between two actions

  The receiver will choose action r_H if her posterior belief exceeds the threshold $\hat{\rho}$;
  choose r_L if her posterior belief is below $\hat{\rho}$;
  may randomize the actions if her posterior belief is $\hat{\rho}$.

- Non-trivial case: receiver will take action r_L without any information, $\rho < \hat{\rho}$

Model

- Sender can send a message $m \in \{m_H, m_L\}$ about his type

  non-verifiable but detectable

- Sender is **lying** if the message is not aligned with his type

  L type sends m_H or H type sends m_L

- A lying cost C > 0

  the sender's intrinsic aversion to lying,

  the potential ex-post penalty for lying,

  the effort of manipulating the information

- Assumption: $C < \min\{\Delta_H^S, \Delta_L^S\}$. Otherwise, sender never lies.

- **Lemma 1**: In equilibrium, type H sender always sends message $m = m_H$.

Model

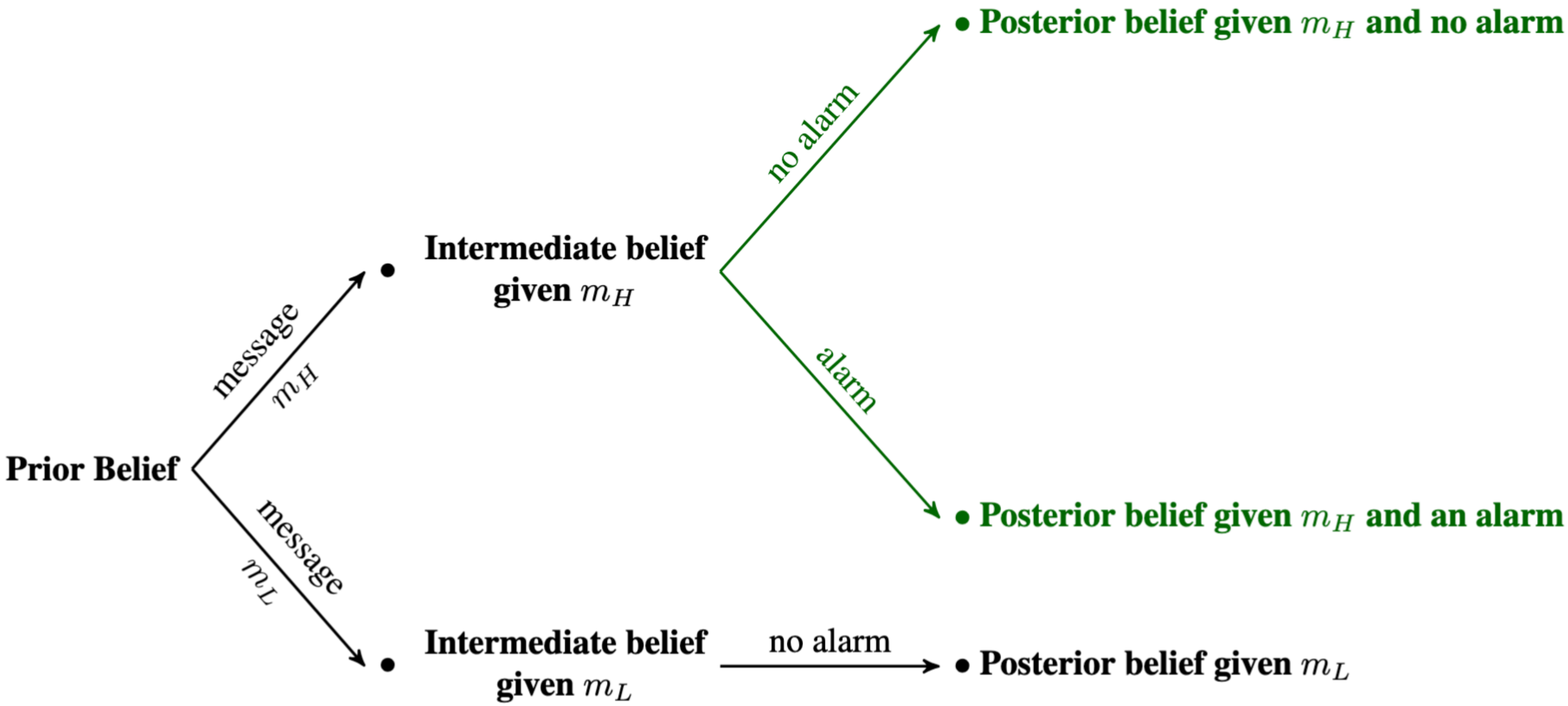
- A lie detector generates a noisy signal $l \in \{a, na\}$ on the truthfulness of the sender's message

  $l = a$ if the message is m_H **and** it thinks the sender is low type

  $l = na$ otherwise

- Receiver infers the sender's type through messages from the sender and the detector.

Receiver's belief updating



Timing

1. The designer designs the lie detector.
2. Nature draws the sender's type $\theta \in \{H, L\}$.
3. The sender sends a message $m \in \{m_H, m_L\}$ to the receiver.
4. The detector sends a signal $l \in \{a, na\}$ to the receiver.
5. The receiver takes an action $r \in \{r_H, r_L\}$.

Detector design

- A designer designs the lie detector.

- Designer's goal depends on the specific contexts

  maximizing the receiver's payoff,

  maximizing the high-type sender's payoff,

  maximizing social welfare/a weighted average of sender and receiver's payoffs

  maximizing more strategic considerations (pricing, platform entry & exit, etc.)

- **True-positive rate** β: probability of sending an alarm when a low-type sender mimics high type, $Pr(l = a \mid m = m_H, \theta = L)$.

- **False-positive rate** α: probability of sending an alarm when the sender is high-type, $Pr(l = a \mid m = m_H, \theta = H)$.

Detector design

- $\{\beta, \alpha\}$: the detector's capacity (quality of detection)

- A stronger detector correctly alarms a lie more frequently and mistakenly alarms a truth-telling message less frequently

Definition 1: A detector $\{\beta', \alpha'\}$ is **stronger** than a detector $\{\beta, \alpha\}$ if and only if the following conditions hold:
$\beta' \geq \beta$, $\alpha' \leq \alpha$, and at least one of the inequalities is strict.

Equilibrium concept

Multi-stage game with incomplete information

↓

Perfect Bayesian Equilibrium (PBE)

Strategies

- Type L sender's strategy

Probability of sending message m_H: σ^S

- Receiver's strategy

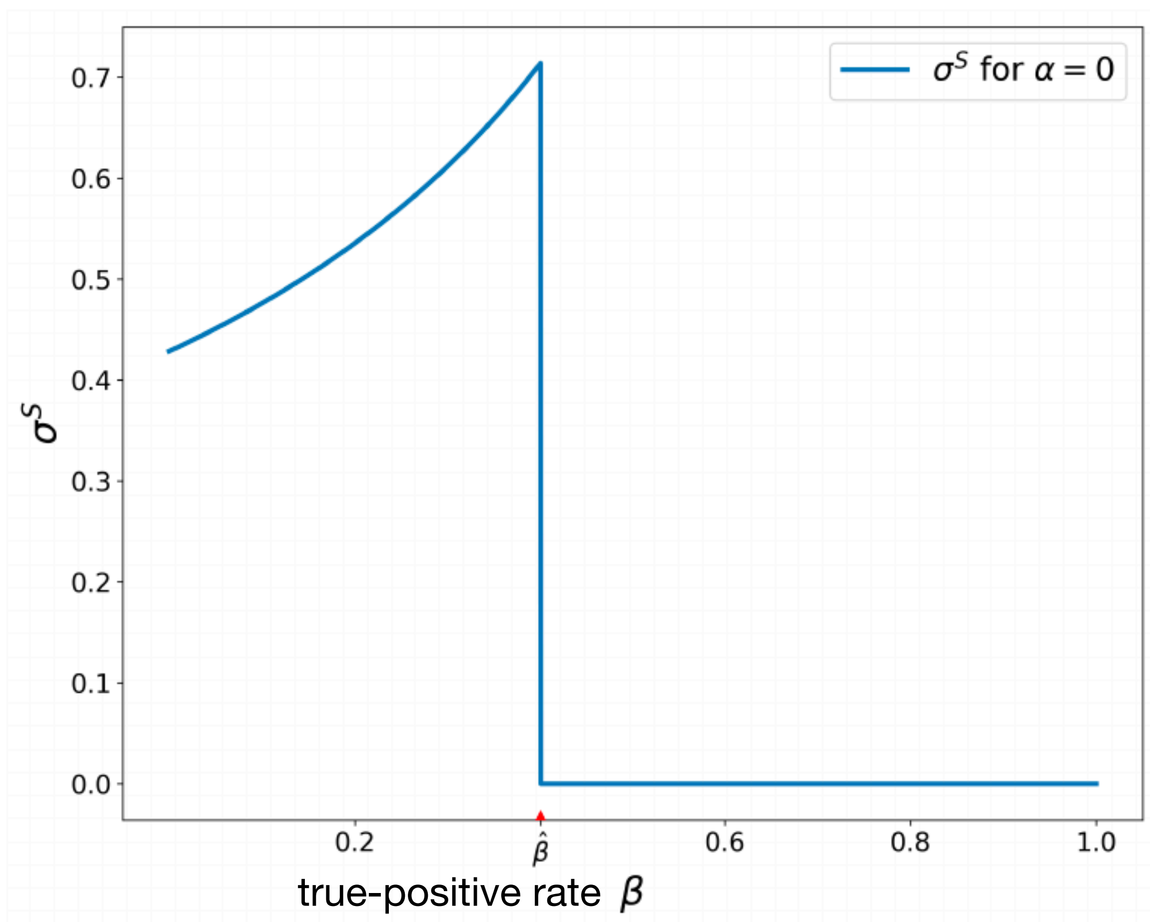
Probability of taking action r_H after seeing $\begin{cases} m_H \text{ and no alarm: } \sigma^R_{na} \\ m_H \text{ and an alarm: } \sigma^R_{a} \\ m_L : \qquad\qquad\quad \sigma^R_{L,na} \end{cases}$

Benchmark

No false-positive alarm benchmark (exogenous detector)

- If a type L sender sends message m_H, the detector sends an alarm with some probability
- False-positive rate $\alpha = 0$, no Type I error

lying probability



$\uparrow$

$$\hat{\beta} = 1 - C/\Delta_L^S$$

No false-positive alarm benchmark (endogenous detector)

Lemma 4 (Endogenous detector)**.** *The receiver's expected payoff, the high-type sender's expected payoff, and the social welfare all (weakly) increase in the true-positive rate β. The optimal true positive rate for the receiver is any $\beta \geq \hat{\beta}$. The optimal true positive rate for the high-type sender is any $\beta \geq \min\left\{1 - \rho\Delta_H^R/[(1-\rho)\Delta_L^R], \hat{\beta}\right\}$. The optimal true positive rate for social welfare is any $\beta \geq \hat{\beta}$.*

- The more accurate the detector is, the better
- The designer always prefers a higher true-positive rate
- No trade-off !

Main Analysis

Equilibrium with an exogenous detector

Equilibrium

The PBEs are:

$$\sigma^{S^*}\begin{cases} = \min\left\{\frac{(1-\alpha)\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R},1\right\}, \beta < \hat{\beta} \\ \in \left[\frac{\alpha\rho\Delta_H^R}{\beta(1-\rho)\Delta_L^R}, \min\left\{\frac{(1-\alpha)\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R},1\right\}\right], \beta = \hat{\beta} \\ = \frac{\alpha\rho\Delta_H^R}{\beta(1-\rho)\Delta_L^R}, \beta > \hat{\beta} \end{cases}$$

$$\sigma_{na}^{R\,*}\begin{cases} = 1, \beta > 1 - \frac{(1-\alpha)\rho\Delta_H^R}{(1-\rho)\Delta_L^R} \\ \in \left[\min\left\{\frac{C}{(1-\beta)\Delta_L^S},1\right\},1\right], \beta = 1 - \frac{(1-\alpha)\rho\Delta_H^R}{(1-\rho)\Delta_L^R} \\ = \min\left\{\frac{C}{(1-\beta)\Delta_L^S},1\right\}, \beta \in (\alpha, 1 - \frac{(1-\alpha)\rho\Delta_H^R}{(1-\rho)\Delta_L^R}) \end{cases}$$

$$\sigma_a^{R^*} = \max\left\{\frac{C}{\beta\Delta_L^S} - \frac{1-\beta}{\beta},0\right\}$$

Effect of lie detection on receiver's posterior belief

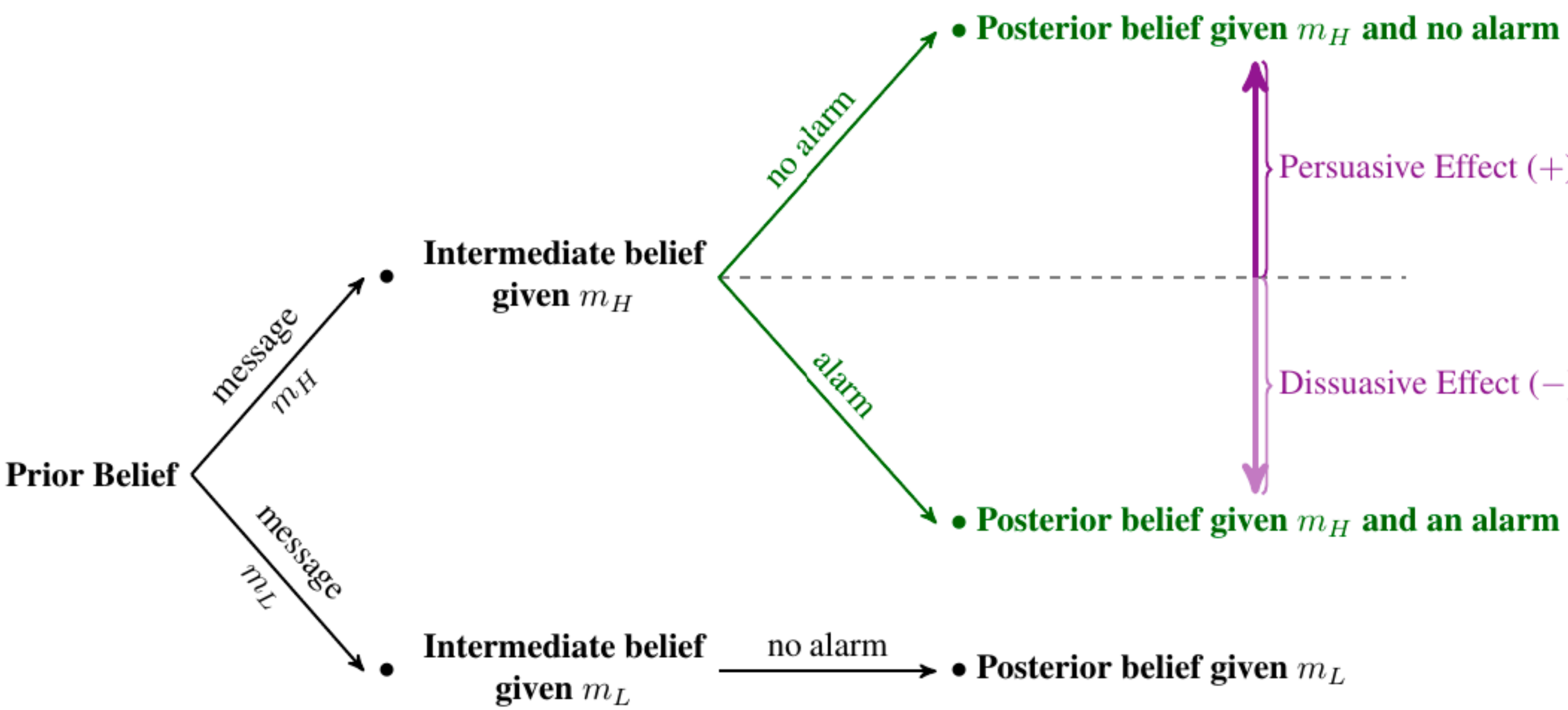


Figure 3: The Effect of Lie Detection on the Receiver's Belief.

Persuasive effect (no alarm)

- The detector is more likely to send **no alarm** when the sender is high-type rather than low-type

- Receiver becomes more certain that the sender is high-type if she receives **no alarm**

  The presence of a detector persuades the receiver to trust the sender's m_H message more

- **Persuasive effect**: the posterior belief-enhancing effect

- The persuasive effect is larger under a stronger detector

Dissuasive effect (alarm)

- The detector is more likely to send an **alarm** when the sender is low-type rather than high-type

- Receiver becomes more certain that the sender is low-type if she receives an **alarm**

  The presence of an alarm makes the receiver less trustful about the sender's m_H message

- **Dissuasive effect**: the posterior belief-reducing effect

- The (absolute value of the) dissuasive effect is larger under a stronger detector

Dissuasive effect (alarm)

- No false-positive alarm benchmark: an alarm eliminates all the uncertainty about the sender's type

⇒ the dissuasive effect does **not** depend on β

Two detectors with very different β generate the **same** effect on the posterior belief

- Main model: two detectors with the same α but different β generate **different** dissuasive effects

⇒ variations in the dissuasive effects lead to qualitatively different equilibrium outcomes

Effect of a stronger detector on the belief

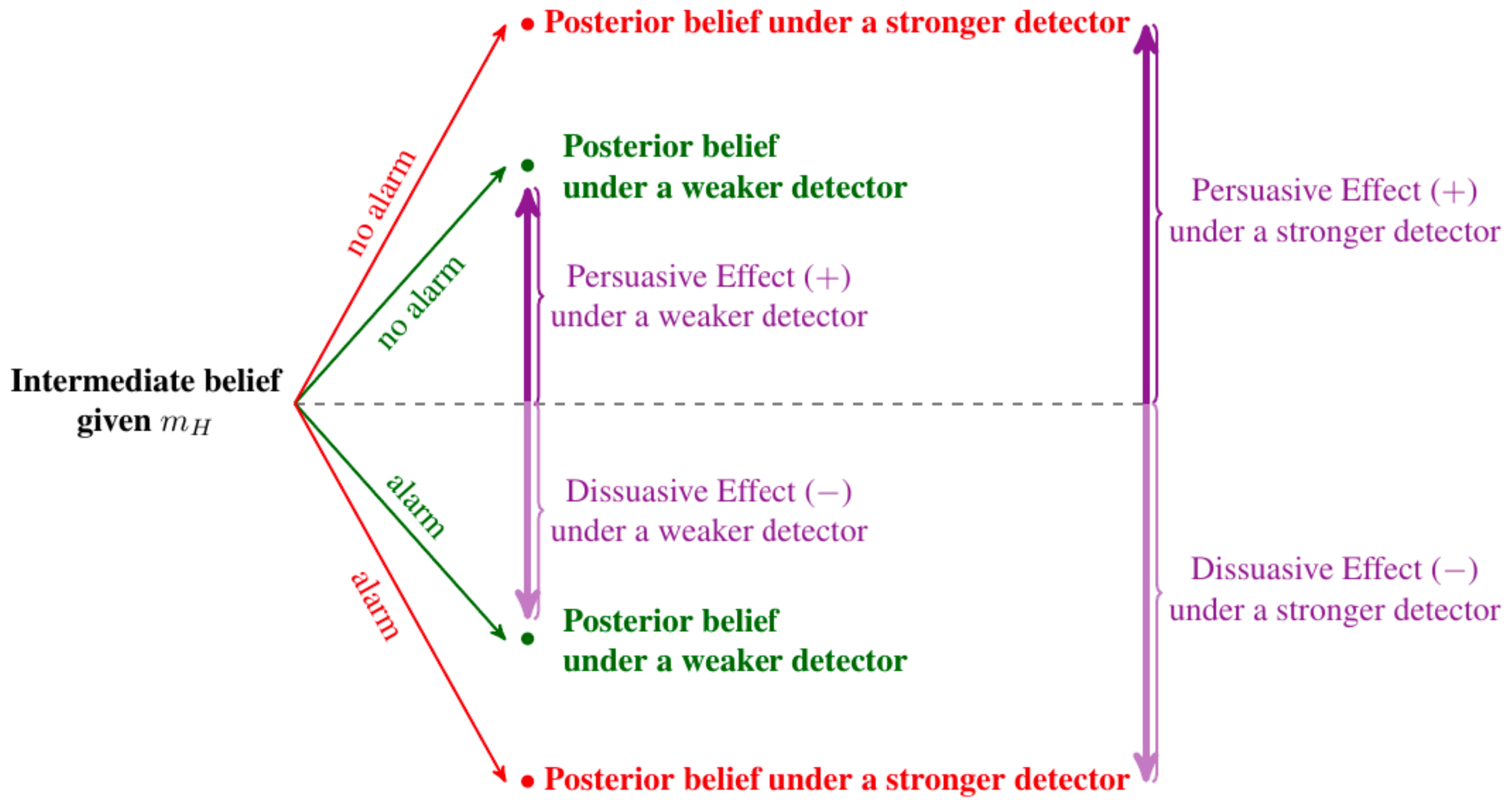
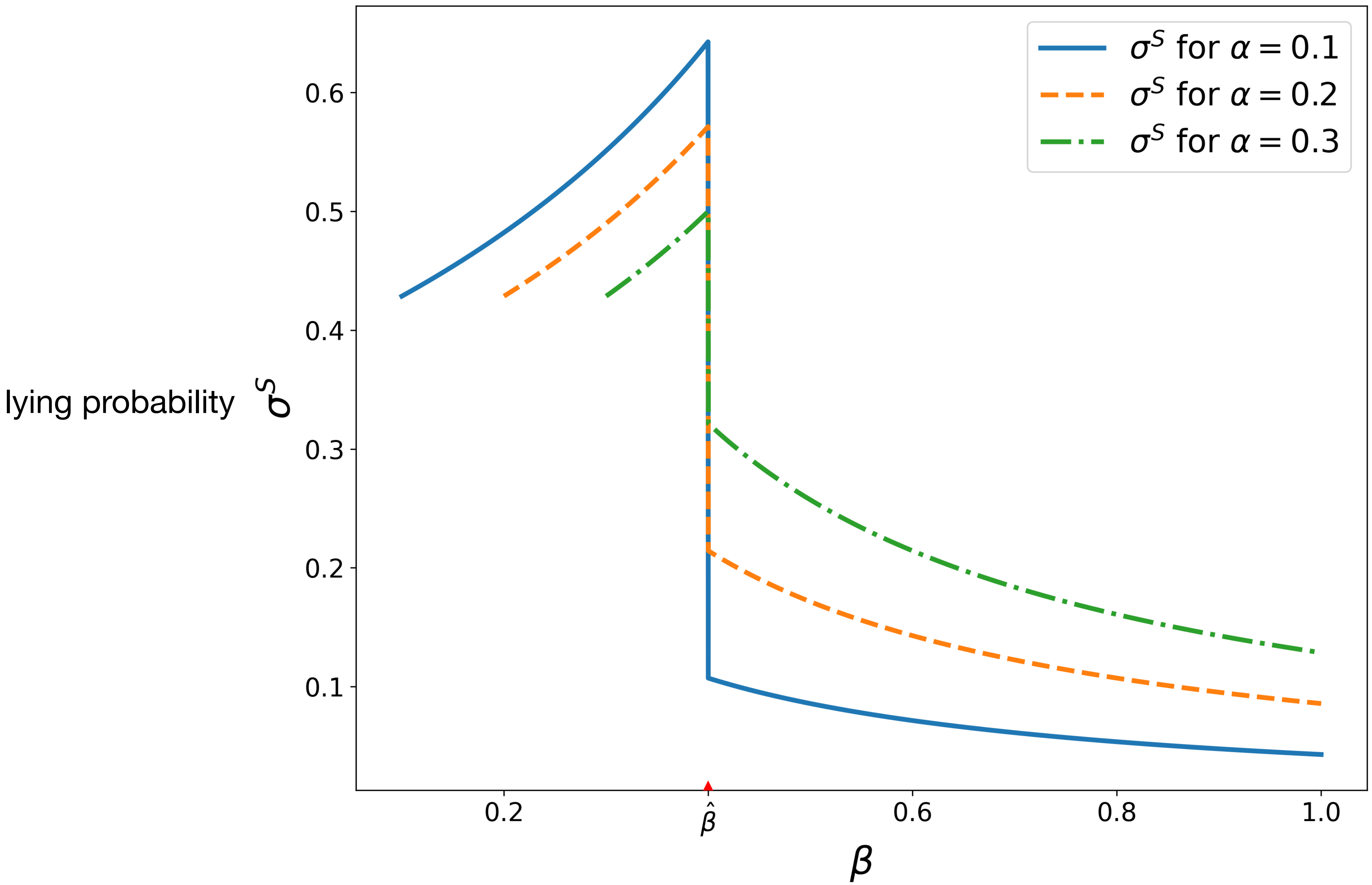
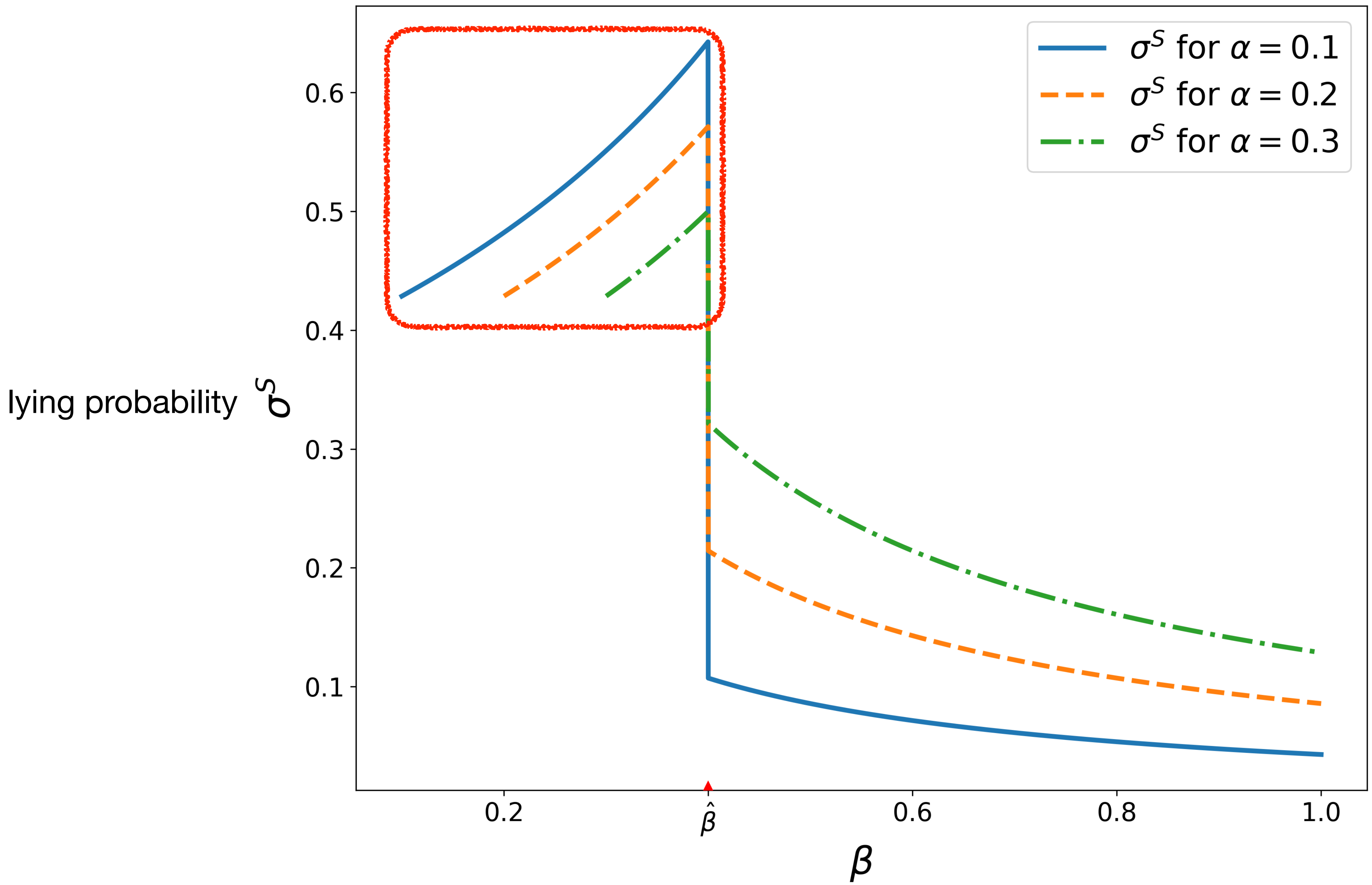


Figure 4: The Effect of a Stronger Lie Detector on the Receiver's Belief.

Non-monotonic relationship between the detector's capacity and the sender's probability of lying



Mechanism (low β)



Intuition on the non-monotonic relationship (low β)

- low $\beta \Rightarrow$ fail to catch many low-type senders who are lying

  $\Rightarrow$ strong incentive for a low type sender to mimic high type

  $\Rightarrow$ high probability of lying

  $\Rightarrow$ low intermediate belief given message m_H

  $\Rightarrow$ low posterior belief ($< \hat{\rho}$ given an alarm)

  $\Rightarrow$ receiver never takes action r_H upon observing an alarm

Intuition on the non-monotonic relationship (low β)

- If the receiver always takes action r_H after m_H and no alarm

Low $\beta \Rightarrow$ high benefit of lying $(1-\beta)\Delta_L^S >$ cost of lying C

$\Rightarrow$ a low-type sender will always lie

$\Rightarrow$ low intermediate belief given message m_H

$\Rightarrow$ low posterior belief also low without an alarm

$\Rightarrow$ receiver should not take action r_H, a contradiction!

Intuition on the non-monotonic relationship (low β)

- If the receiver never takes action r_H after m_H and no alarm

$\Rightarrow$ no low-type sender will lie

$\Rightarrow$ only high-type senders will send message m_H

$\Rightarrow$ posterior belief = 1 given message m_H (regardless of the alarm)

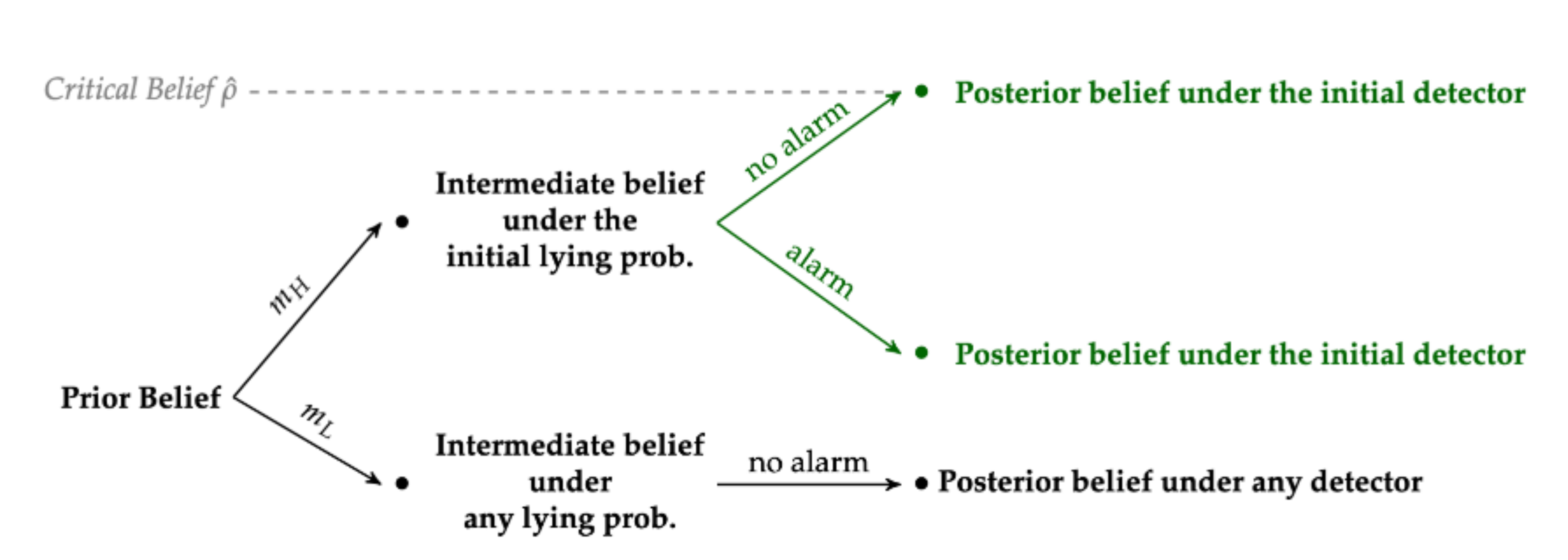
$\Rightarrow$ receiver should always take action r_H upon observing m_H, a contradiction!

Intuition on the non-monotonic relationship (low β)

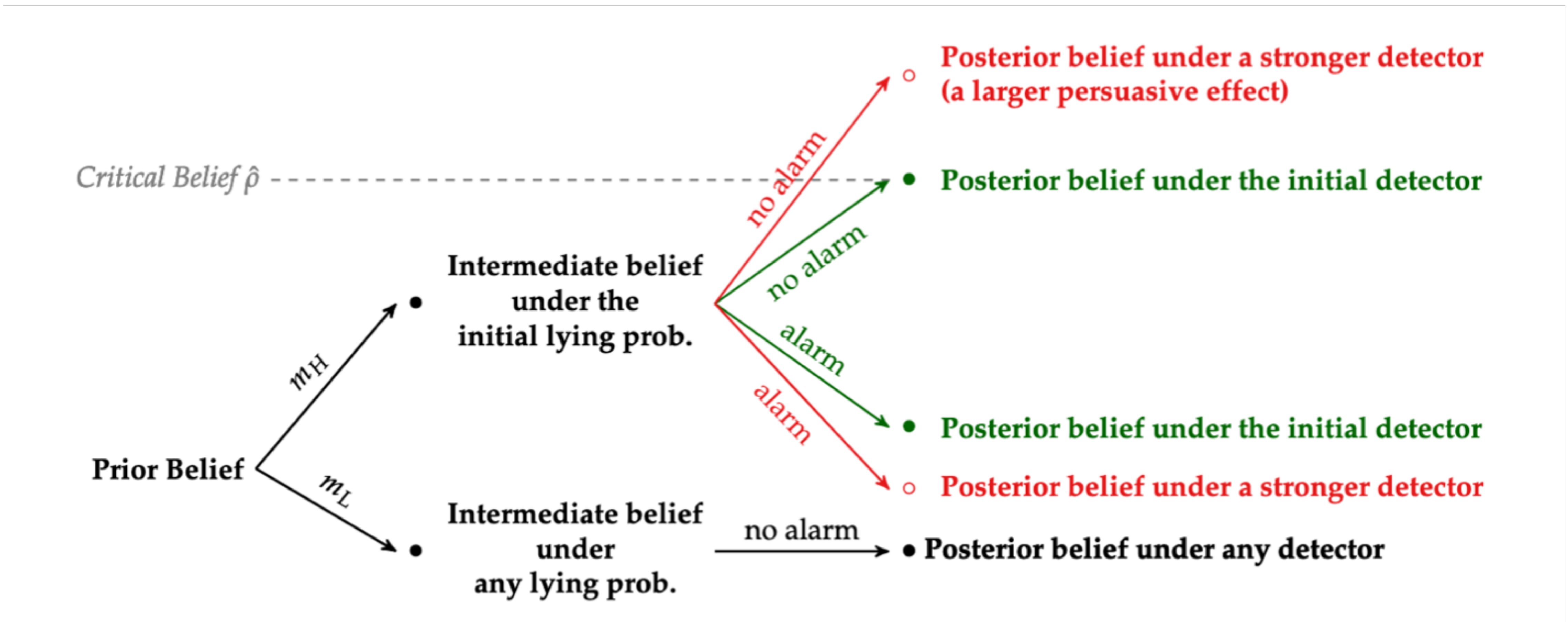
- The receiver must use a mixed strategy after m_H and no alarm

$\Rightarrow$ posterior belief $= \hat{\rho}$ given m_H and no alarm

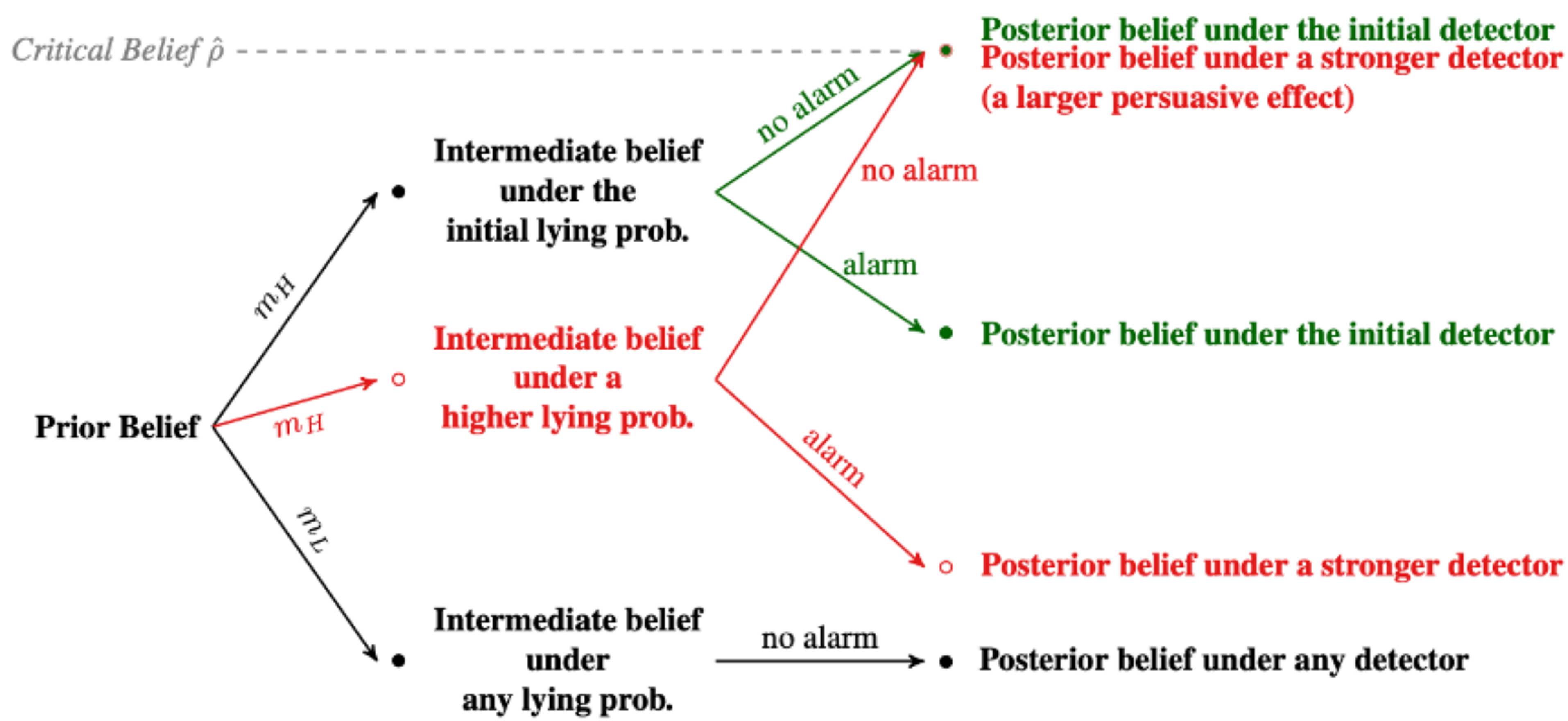
Intuition on the non-monotonic relationship (low β)



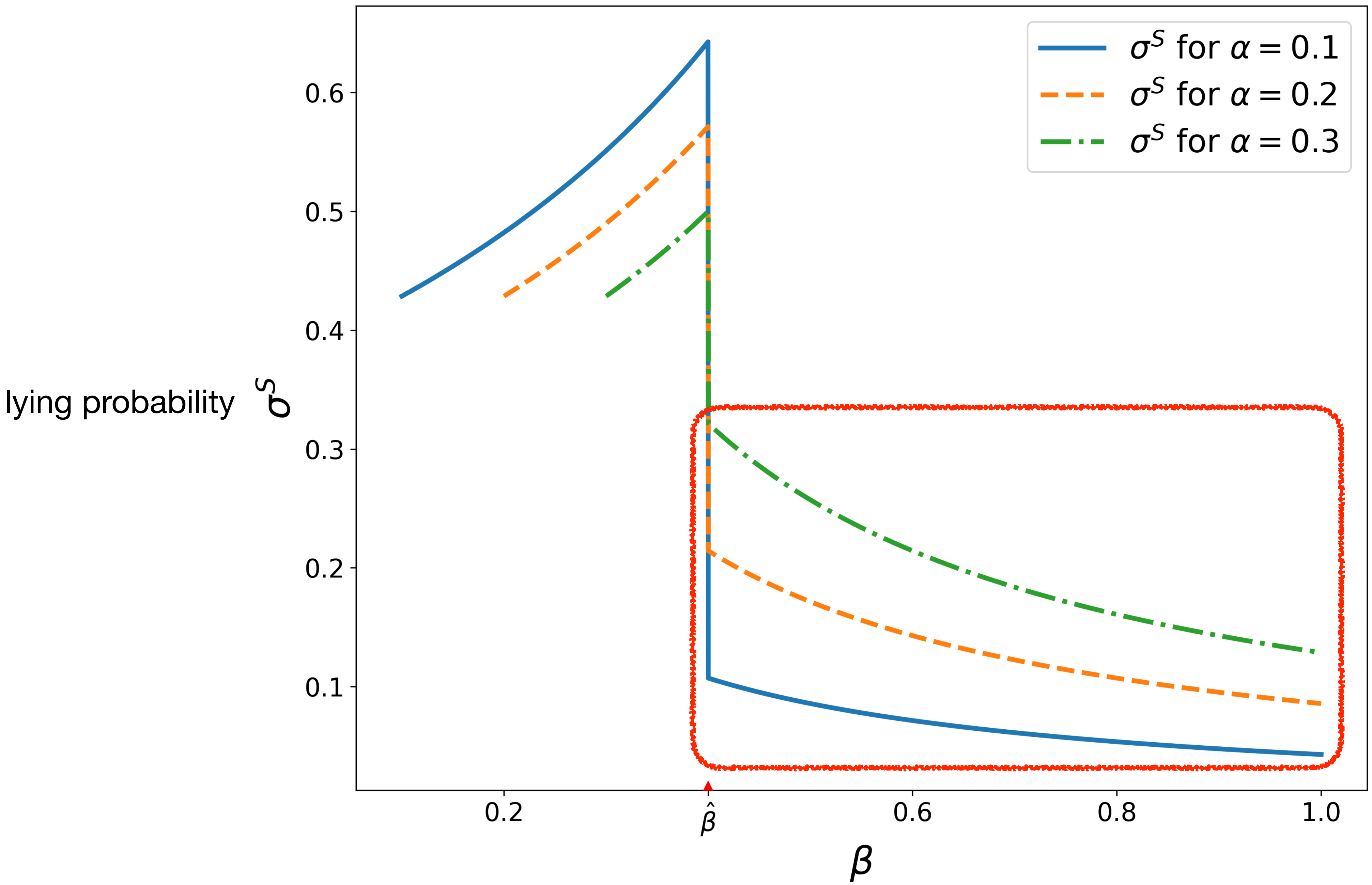
Intuition on the non-monotonic relationship (low β)



Intuition on the non-monotonic relationship (low β)



Mechanism (high β)



Intuition on the non-monotonic relationship (high β)

- High $\beta \Rightarrow$ catch a high proportion of low-type senders who are lying

  $\Rightarrow$ low incentive for a low type sender to mimic high type

  $\Rightarrow$ low probability of lying

  $\Rightarrow$ high intermediate belief given message m_H

  $\Rightarrow$ high posterior belief ($> \hat{\rho}$ without an alarm)

  $\Rightarrow$ receiver always takes action r_H when there is no alarm

Intuition on the non-monotonic relationship (high β)

- If the receiver always takes action r_H after m_H and an alarm

  High benefit of lying $\Delta_L^S >$ cost of lying C

  $\Rightarrow$ a low-type sender will always lie (pooling equilibrium)

  $\Rightarrow$ intermediate belief = prior belief < $\hat{\rho}$

  $\Rightarrow$ posterior belief given an alarm < intermediate belief < $\hat{\rho}$

  $\Rightarrow$ receiver should not take action r_H, a contradiction!

Intuition on the non-monotonic relationship (high β)

- If the receiver never takes action r_H after m_H and an alarm

High $\beta \Rightarrow$ low benefit of lying $(1-\beta)\Delta_L^S$ < cost of lying C

$\Rightarrow$ no low-type sender will lie

$\Rightarrow$ only high-type senders will send message m_H

$\Rightarrow$ posterior belief = 1 given message m_H (regardless of the alarm)

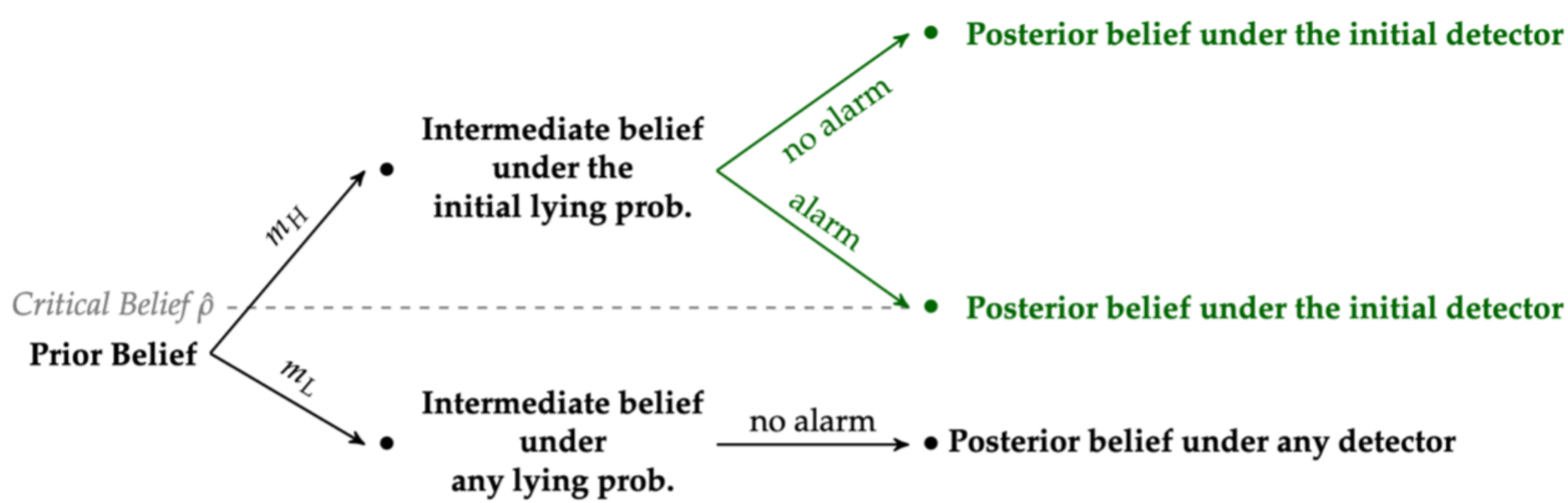
$\Rightarrow$ receiver should always take action r_H upon observing m_H, a contradiction!

Intuition on the non-monotonic relationship (high β)

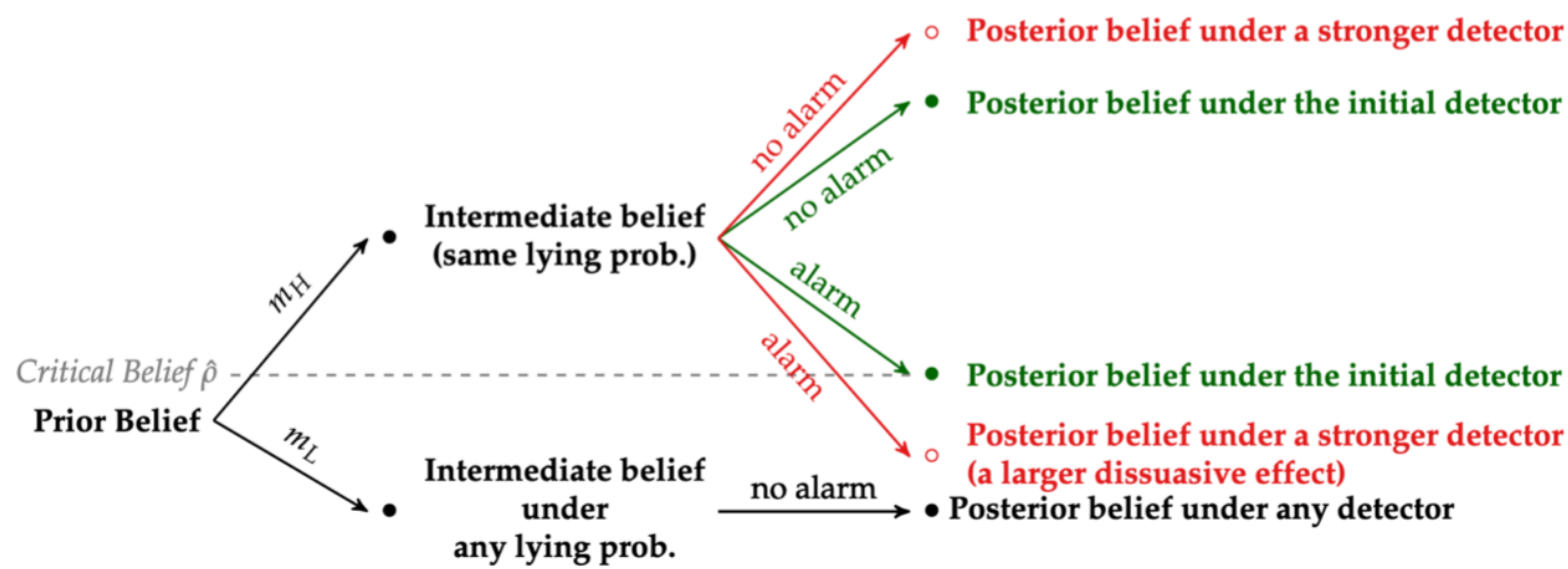
- The receiver must use a mixed strategy after m_H and an alarm

  $\Rightarrow$ posterior belief $= \hat{\rho}$ given m_H and an alarm

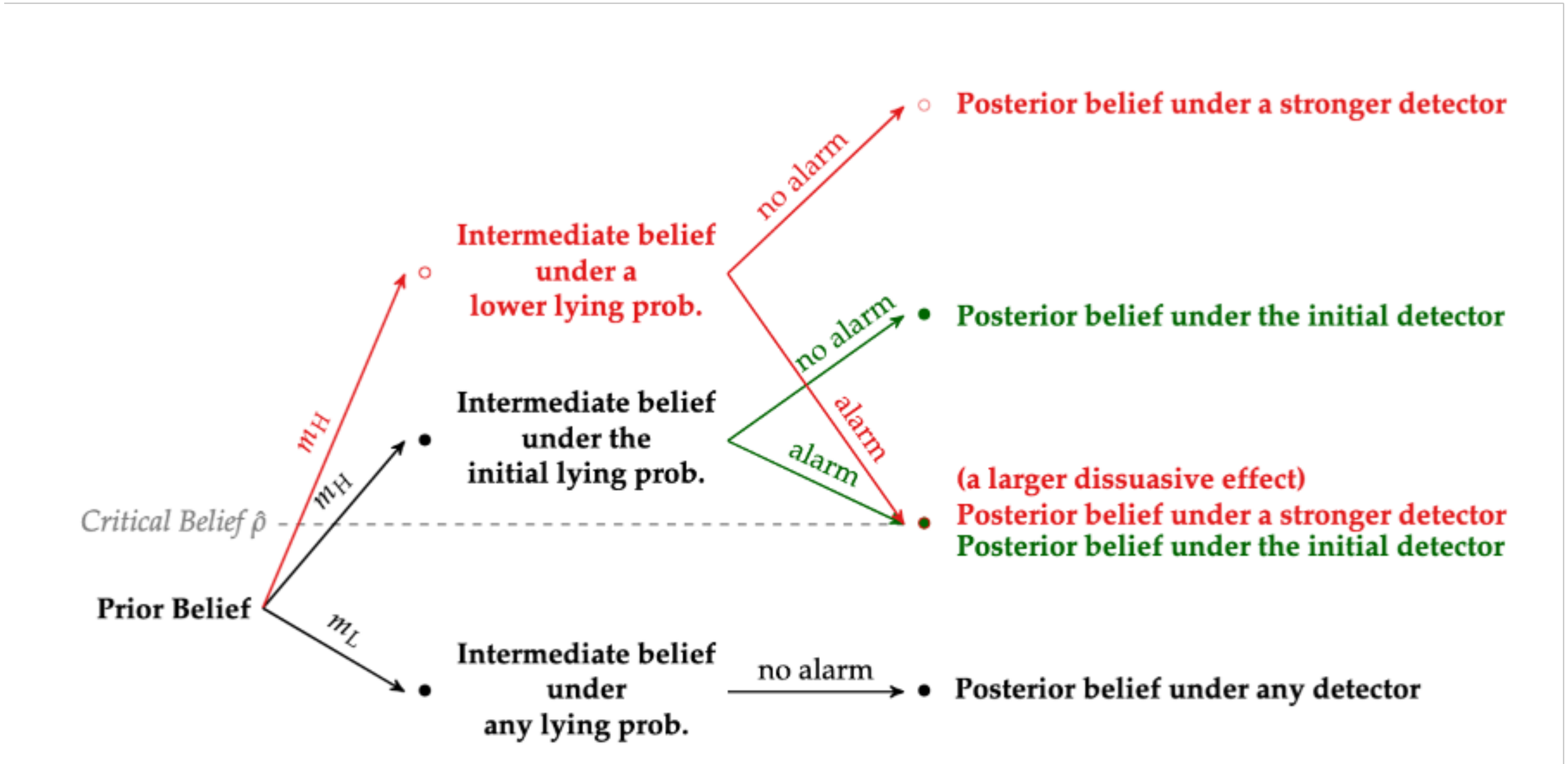
Intuition on the non-monotonic relationship (high β)



Intuition on the non-monotonic relationship (high β)

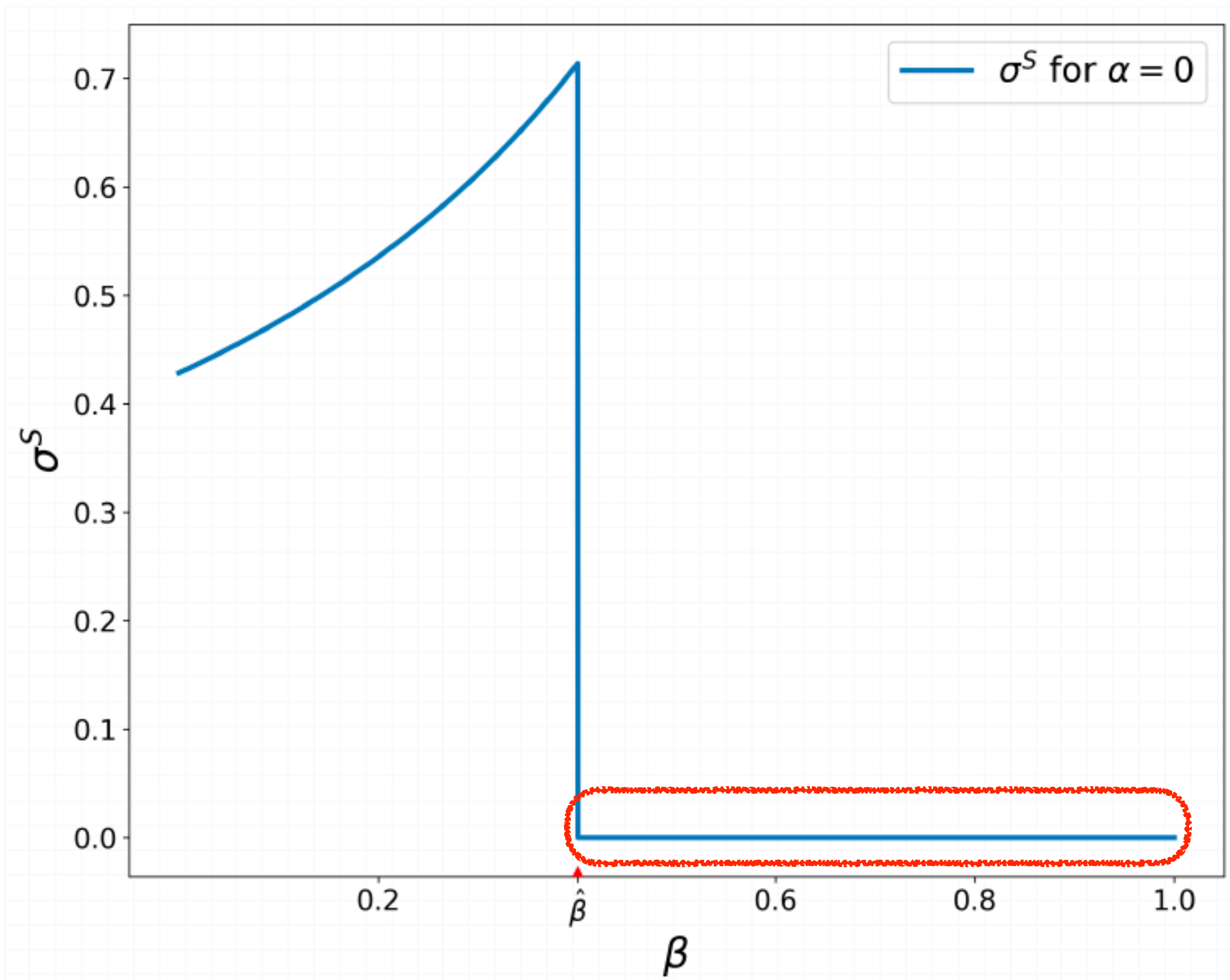


Intuition on the non-monotonic relationship (high β)



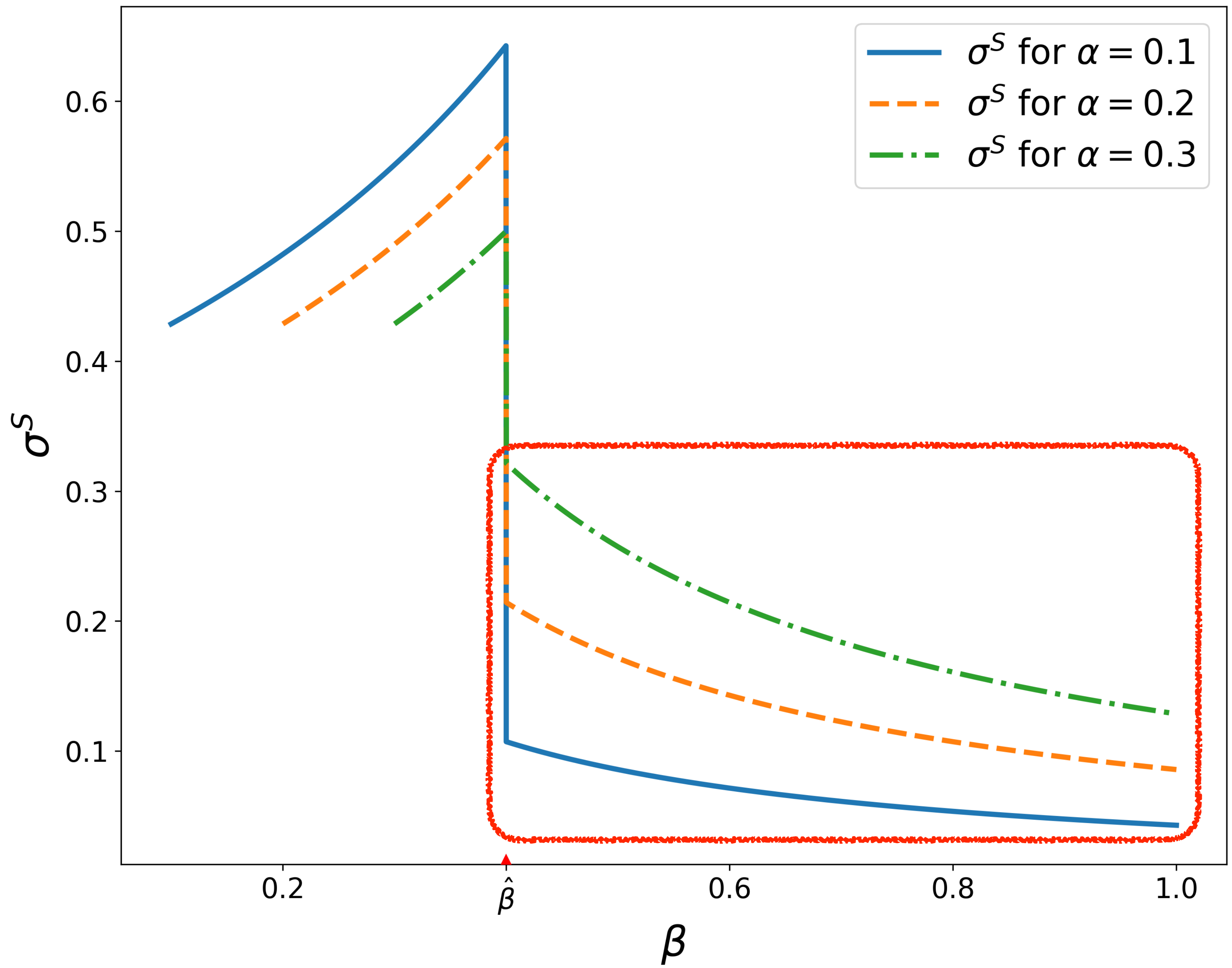
No-false-positive benchmark: no disinformation under high β

lying probability



Main model: disinformation always exists

lying probability



Main model: disinformation always exists

- Suppose the sender never lies

  ⇒ only high-type senders will send message m_H

  ⇒ an alarm must be a false-positive alarm

  ⇒ by deviating, a low-type sender will never be caught

  ⇒ low-type senders will lie

  ⇒ no-lying (separating) equilibrium cannot be sustained

Entire Equilibrium
(endogenous detector)

Detector design

- Designer has access to an exogenously given classifier that generates a prediction for the message's trustworthiness (technology constraint)
- Designer decides whether to send an alarm based on the prediction
- The classifier generates a binary outcome $s \in \{s_L, s_H\}$
- $\phi(s \mid \theta)$: probability of outcome $s \in \{s_L, s_H\}$ conditional on the sender's true type $\theta \in \{\theta_L, \theta_H\}$
- ϕ: the classifier's capacity (quality of classification)

  Perfectly informative: $\phi(s_H \mid \theta = H) = \phi(s_L \mid \theta = L) = 1$

  Not informative: $\phi(s_H \mid \theta = H) \approx \phi(s_H \mid \theta = L)$ & $\phi(s_L \mid \theta = L) \approx \phi(s_L \mid \theta = H)$

Detector design

- Designer's decision:

  prob. of sending an alarm given classification outcome s_L, $\lambda_L = \Pr(l = a \mid s_L)$

  prob. of sending an alarm given classification outcome s_H, $\lambda_H = \Pr(l = a \mid s_H)$

- $\{\lambda_L, \lambda_H\}$: the alarm rule

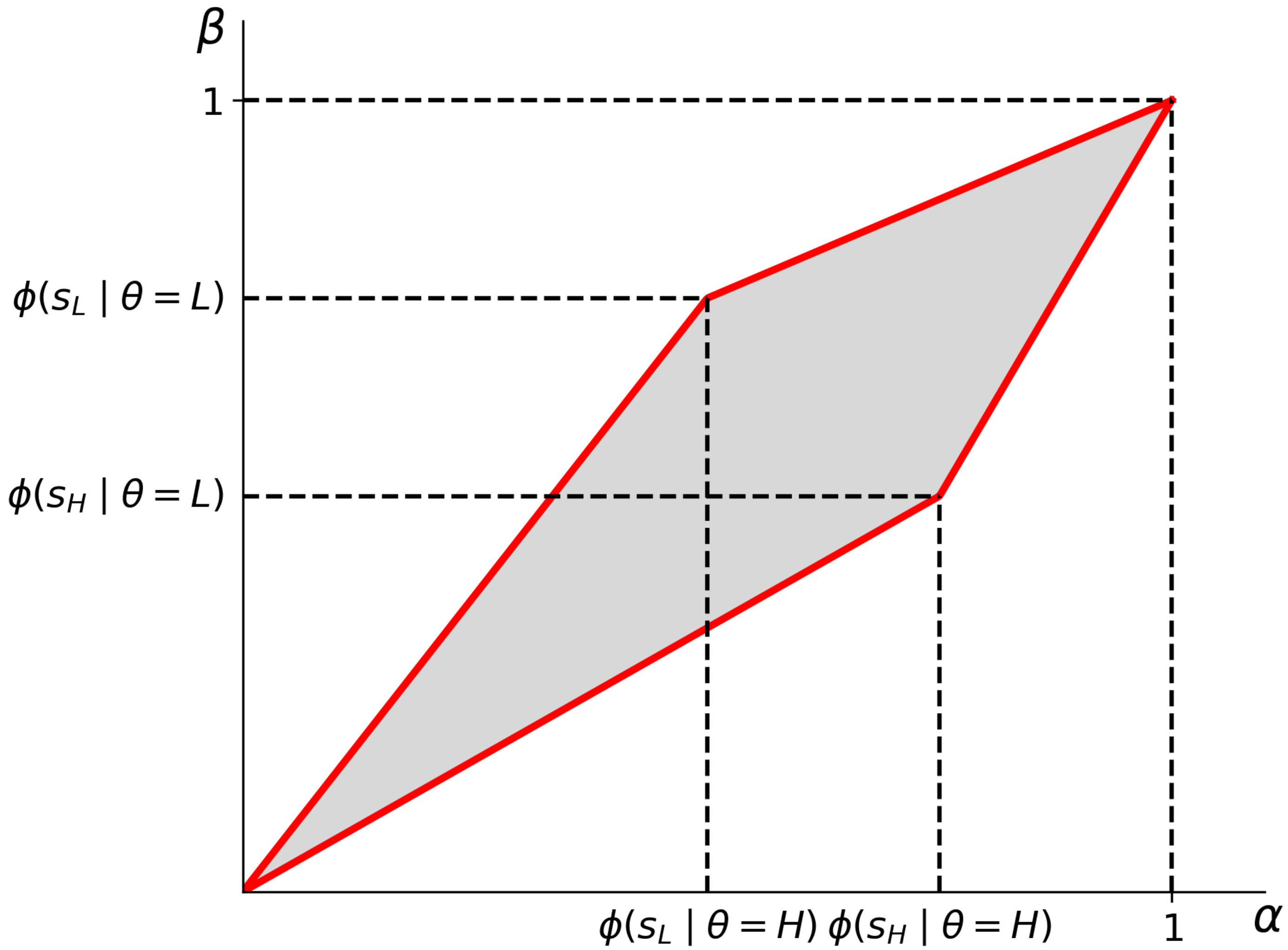
- Classifier's capacity + alarm rule $\rightarrow$ detector's capacity (β, α)

Feasible detector space

- The designer cannot obtain all detectors $(\beta, \alpha) \in \{(\beta, \alpha) \mid 0 \le \alpha < \beta \le 1\}$ due to the constraint of the classifier's capacity

- Space of the feasible detectors given a classifier ϕ:
$$\{(\beta, \alpha) \mid \beta = \phi(s_L \mid \theta = L)\lambda_L + \phi(s_H \mid \theta = L)\lambda_H,$$
$$\alpha = \phi(s_L \mid \theta = H)\lambda_L + \phi(s_H \mid \theta = H)\lambda_H,$$
$$\lambda_L \in [0,1], \lambda_H \in [0,1]\}$$

Feasible detector space



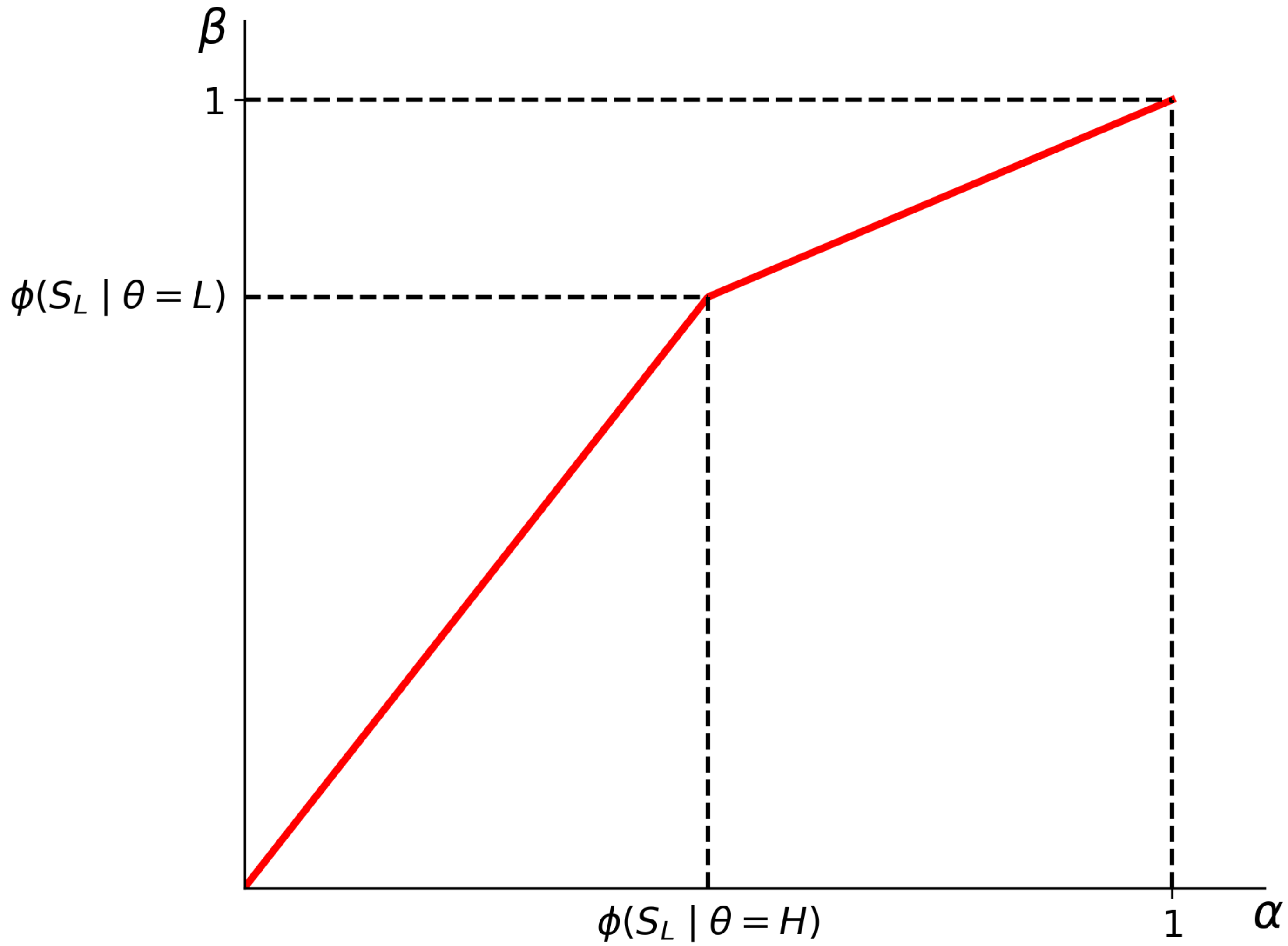
Effect of lie detection on payoffs

- For a given α, a higher β benefits the receiver and high-type sender, whereas hurts the low-type sender

- For a given β, a lower α makes all players better off

$\Rightarrow$ The designer always chooses the **lowest feasible false-positive rate** α given any true-positive rate β.

Receiver operating characteristic (ROC) curve



- Pareto frontier of the classification outcome
- the lowest feasible false-positive rate α given any true-positive rate β

Optimal false-positive rate and alarm Rule given true-positive rate

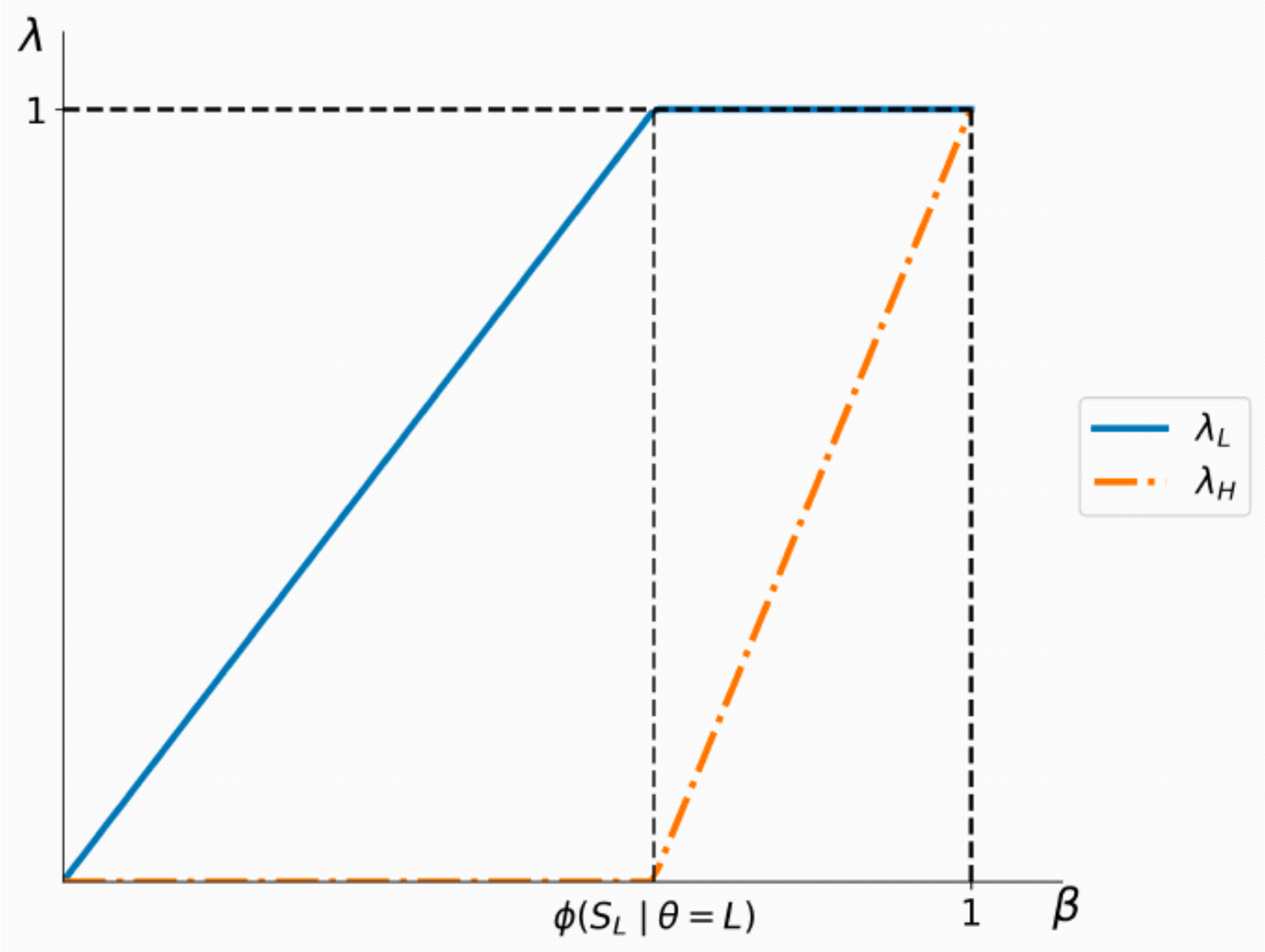
Lemma 5 (Optimal False-positive Rate and Alarm Rule Given True-positive Rate)**.** *For a given true-positive rate β, the detector's optimal false-positive rate, denoted by $\alpha^*(\beta;\phi)$, is*

$$\alpha^*(\beta;\phi) = \begin{cases} \frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\beta, & if\ \beta \leq \phi(s_L|\theta=L) \\ \frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}\beta + 1 - \frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}, & if\ \beta > \phi(s_L|\theta=L), \end{cases}$$

which increases in β. The detector $\{\beta, \alpha^*(\beta;\phi)\}$ can be achieved by the alarm rule

$$\lambda_L^*(\beta) = \begin{cases} \frac{\beta}{\phi(s_L|\theta=L)}, & if\ \beta \leq \phi(s_L|\theta=L) \\ 1, & if\ \beta > \phi(s_L|\theta=L) \end{cases}, \ \lambda_H^*(\beta) = \begin{cases} 0, & if\ \beta \leq \phi(s_L|\theta=L) \\ \frac{\beta-\phi(s_L|\theta=L)}{\phi(s_H|\theta=L)}, & if\ \beta > \phi(s_L|\theta=L). \end{cases}$$

Optimal alarm rule given a true-positive rate



- The designer wants to achieve a given true-positive rate while minimizing the false-positive rate.

Optimal Design of Detector

- Optimal choice of a feasible detector, $\{\beta^*, \alpha^*\}$

- Previous lemma: optimal false-positive rate is $\alpha^*(\beta; \phi)$ given any true-positive rate β.

$\Rightarrow$ only need to pin down the optimal true-positive rate β^*

- **Lemma 6:** Sequential derivation $\Leftrightarrow$ simultaneous optimization

High/low capacity of a classifier

- The classifier can better distinguish the sender's type if it generates s_H more frequently when the sender is high type than low type (higher $\phi(s_H | \theta = H)/\phi(s_H | \theta = L)$);

- It also can better distinguish the sender's type if it generates s_L more frequently when the sender is low type than high type (higher $\phi(s_L | \theta = L)/\phi(s_L | \theta = H)$).

Definition 2. A classifier has a high capacity if $\phi(s_H|\theta = H)/\phi(s_H|\theta = L) \geq (1-\rho)\Delta_L^R/(\rho\Delta_H^R)$ and $\phi(s_L|\theta = L)/\phi(s_L|\theta = H) \geq (\Delta_L^S - C)\rho\Delta_H^R/[\Delta_L^S\rho\Delta_H^R - (1-\rho)\Delta_L^R C]$. Otherwise, it has a low capacity.

Maximizing receiver's payoff

- **<u>Proposition 3:</u>** The β^* that maximizes the receiver's expected payoff is:

  (1) If the classifier has a low capacity:

  any $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L \mid \theta = L)\}]$, which minimizes the low-type sender's equilibrium probability of lying,

  (2) If the classifier has a high capacity:

  any $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L \mid \theta = L)\}]$ if the lying cost is high;

  $\beta = \phi(s_L \mid \theta = L)$, if the lying cost is low.

- Receiver benefits from a low percentage of disinformation and a good detection technology

$\Rightarrow$ β^* either minimizes the low-type sender's probability of lying or takes full advantage of the region where a unit increase in β leads to a small increase in α

Maximizing high-type sender's payoff

- **Proposition 4:** The β^* that maximizes the high-type sender's expected payoff is:

(1) If the classifier has a low capacity: any $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L | \theta = L)\}]$;

(2) If the classifier has a high capacity:
$\beta_1 := [\rho\Delta_H^R - (1-\rho)\Delta_L^R]/[\rho\Delta_H^R\phi(s_L | \theta = H)/\phi(s_L | \theta = L) - (1-\rho)\Delta_L^R] < \hat{\beta}$, which is decreasg in $\phi(s_L | \theta = L)/\phi(s_L | \theta = H)$,.

- The sender prefers the receiver to always take action r_H conditional on message m_H and no alarm (large enough β)

- Fixing a β, the detector has a lower α if the classifier has a higher capacity.

- Higher posterior belief ⇒ detector can induce action r_H even if β is adjusted downward

- When the classifier has a high capacity, *counter-intuitively*, the optimal detector alarms a **smaller** percentage of disinformation when its underlying classifier is **better** at distinguishing the sender's type.

Maximizing

$$\mathbb{E}U^R(\beta) + w_H \rho \mathbb{E}U_H^S(\beta) + w_L(1-\rho)\mathbb{E}U_L^S(\beta)$$

Proposition 5. *If the classifier has a low capacity, any* $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}]$ *is optimal. If the classifier has a high capacity, the set of optimal true-positive rates is:*

$$\begin{cases} \mathcal{B}(w_H, w_L) & \textit{if } C < \tilde{C}(w_H, w_L) \\ \mathcal{B}(w_H, w_L) \cup [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}] & \textit{if } C = \tilde{C}(w_H, w_L) \\ [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}] & \textit{if } C \in (\tilde{C}(w_H, w_L), \Delta_L^S(1-\beta_1)] \end{cases},$$

where

$$\mathcal{B}(w_H, w_L) := \begin{cases} \{\phi(s_L \mid \theta = L)\} & \textit{if } n_0 w_H + l_0 w_L < 1 \\ [\beta_1, \phi(s_L \mid \theta = L)] & \textit{if } n_0 w_H + l_0 w_L = 1 \\ \{\beta_1\} & \textit{if } n_0 w_H + l_0 w_L > 1 \end{cases},$$

$\tilde{C}(w_H, w_L)$ *is a continuous function increasing in both* w_H *and* w_L*, and* n_0 *and* l_0 *are strictly positive constants.*

Maximizing

$$\mathbb{E}U^R(\beta) + w_H\rho\mathbb{E}U_H^S(\beta) + w_L(1-\rho)\mathbb{E}U_L^S(\beta)$$

- If $w_H = w_L = 1$: maximizing social welfare

- Low-capacity classifier: senders' and receivers' incentives are aligned

- High-capacity classifier: the sender prefers a lower β than the receiver

$\Rightarrow$ optimal detector is a compromise between their preferences

$\Rightarrow$ greater weight on senders, $\uparrow w_H, w_L \rightarrow$ lower β*

Comparison with the no false-positive alarm benchmark

- No false-positive alarms: no trade-off in the detector design $\Rightarrow$ designer always prefers a higher β.

- With false-positive alarms: designer strictly prefers intermediate β.

False-positive alarms + players' strategic responses $\rightarrow$ a higher β may reduce the receiver's payoff, the high-type sender's payoff, and social welfare

Extensions

Extensions

- Restriction on the alarm rule: $\lambda_H = 0$
- Endogenous commission fee and detector design

Thanks!